

TOWARDS SIGN LANGUAGE RECOGNITION BASED ON BODY PARTS RELATIONS

M. Martinez-Camarena

Universidad Politecnica de Valencia

J. Oramas M., T. Tuytelaars

KU Leuven, ESAT-PSI, iMinds

ABSTRACT

Over the years, hand gesture recognition has been mostly addressed considering hand trajectories in isolation. However, in most sign languages, hand gestures are defined on a particular context (body region). We propose a pipeline which models hand movements in the context of other parts of the body captured in the 3D space using the Kinect sensor. In addition, we perform sign recognition based on the different hand postures that occur during a sign. Our experiments show that considering different body parts brings improved performance when compared with methods which only consider global hand trajectories. Finally, we demonstrate that the combination of hand postures features with hand gestures features helps to improve the prediction of a given sign.

1. INTRODUCTION

There is a wide variety of sign languages that are used by hearing-impaired individuals. Each language is formed by grammar rules and a vocabulary of signs. Something that most of these languages have in common is that signs are composed by two elements: *hand postures*, i.e. the position or configuration of the fingers; and *hand gestures*, i.e. the movement of the hand as a whole (see Fig. 1). In this paper we focus on the problem of sign classification based on hand postures and hand gestures, leaving elements such as facial gestures or grammar rules for future work.

In this work, we consider relations between different parts of the body for the task of sign language recognition. For example, see how the global motion of the sign in Fig. 1(a) is very similar to the motion of the sign in Fig. 1(b). However, the relative motion of the hand (magenta) w.r.t. the head (green) is different for both signs, especially at the very end. Our method uses a MS Kinect to capture RGBD images, as well as to localize the different body parts [1]. Then, we represent each sign by a combination of responses obtained from cues extracted from hand postures and hand gestures, respectively. For the problem of sign language recognition based on hand posture cues, we use shape context descriptors [2] in combination with a multiclass SVM classifier [3] to recognize the different signs. Regarding sign recognition based on cues derived from hand gestures, we use Hidden Markov Models (HMMs) to model the dynamics of each gesture. Finally, for the task of sign language recognition, we

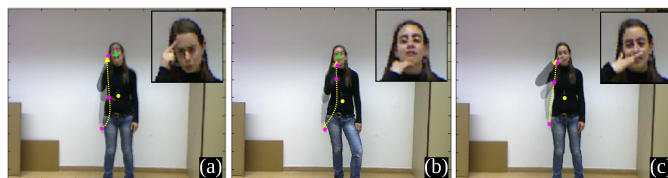


Fig. 1: Note how signs with similar global trajectories, (a) and (b), can be distinguished based on the relative locations of the hand w.r.t. the head. In addition, see how the posture of the hands can help to distinguish between similar signs. Selected body part locations in color. Green: head location, magenta: right hand, yellow: torso (Best viewed in color).

compute the sign prediction by late fusion of the responses of the processes for sign recognition based on hand postures and gestures, respectively, via multiclass SVM classification.

The main contribution of this work is to show that reasoning about relations between parts of the body brings improvements for hand gestures recognition and has potential for sign language recognition. This paper is organized as follows: Section 2 positions our work with respect to similar work. In Section 3 we present the details of our method. Section 4 presents the evaluation protocol and experimental results. In Section 5 we conclude this paper.

2. RELATED WORK

Regarding hand gesture recognition, the skeleton-based algorithms make use of 3D information to identify key elements, in particular the body parts. Shotton et al. [1] presented a milestone method for the extraction of the human body skeleton. The approach from [1] allows relatively accurate tracking of the parts of the body in real-time. The method from [4] converts the set of skeleton joints in each frame to joint angles and their respective angular velocities representation. Based on these descriptors, their method is able to identify actions such as: clapping, throwing, punching, etc. In [5] Chai et al. follow the 3D trajectory of the hand joint normalizing by linear re-sampling. Then, they perform a trajectory alignment to compare, based on matching scores, with a set of known trajectories. Similar to these works, we use an implementation of the algorithm from [1] to acquire the set of points of the skeleton in each frame. In addition, we build a descriptor modeling the joints of the hands w.r.t. the joints of the other parts of the body.

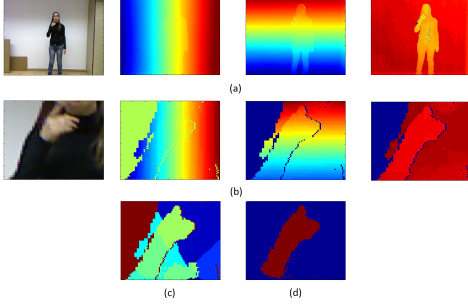


Fig. 2: Hand segmentation algorithm: (a) RGB image collected with kinect and the computed 3D points (X,Y,Z) projected in the image space, (b) RGB image and projected 3D points after spatial thresholding, (c) 3D points assigned to the different joints of the body, and (d) Binary hand region H .

3. PROPOSED METHOD

The proposed system can be summarized in the following steps: MS Kinect is used to capture the RGBD images. We estimate the skeleton representation from these images using the algorithm from [1]. Then, our system consists of 2 parallel stages: the recognition of signs based on hand posture features and the recognition of signs based on hand gesture features. Finally, these responses are combined to estimate the likelihood of a given sign.

3.1. Sign Recognition based on Hand Postures

Hand Region Segmentation: The component based on hand posture features takes as input RGBD images and the skeleton representation estimated using kinect and the algorithm from [1]. Then, we compute the 3D coordinates (X,Y,Z) of all the points of the scene from the depth images (Fig. 2(a)). To reduce the number of points to be processed, we perform an early spatial threshold to ignore points far from the expected hand regions. To this end, we remove all the points outside the sphere centered on the hand joint whose radius is half the distance between the joints of the hand and elbow (Fig. 2(b)). Then, we assign the remaining 3D points to the closest body joint, using Nearest Neighbors (NN) classification (Fig. 2(c)). This assignment is computed in the 3D space, keeping correspondences with the pixels in the image space. Following the assignment, we only keep the points assigned to the joints of the hands. For the case of multiple regions assigned to the hand joint, we keep the largest region. This allows our method to overcome noise introduced by low resolution images and scenarios in which the hand comes in contact with other parts of the body. In addition, we re-scale the depth images to a 65x65 pixels patch. Finally, as Fig. 2(d) shows, we binarize the re-scaled patch producing the hand region H .

Hand Posture Description: Once we have obtained the 2D regions H containing the hands, we describe the different hand postures by a Bag-of-Words representation constructed from shape context descriptors [2]. In order to compute the shape

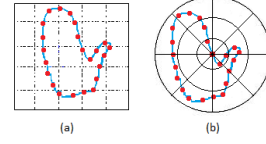


Fig. 3: Computation of Shape Context descriptors: (a) Selection of equally-spaced points on the hand region H contour, and (b) Log polar sampling (8 angular and 3 distance bins).

context descriptor s , we extract a number of m equally-spaced points from the contour of each hand region H (Fig. 3(a)) obtained from the hand segmentation step. Then, using this set of points, a log-polar binning coordinate system is centered at each of the points and a histogram accumulates the amount of contour points that fall within each bin (Fig. 3(b)). This histogram is the shape context descriptor. This procedure is performed on each frame of the videos. Then, we define a bag-of-words representation p where each video is a bag containing a set of words from a dictionary obtained by vector-quantizing the shape context descriptors s via K-means. In our experiments we use $K = 100$ since that value gave the best performance in the validation set. This procedure is applied for both hands of the user producing two descriptors, (p_{right}, p_{left}) , one for each hand, which are concatenated into one posture-based descriptor $p = [p_{right}, p_{left}]$.

Recognizing signs based on hand postures features: Once we have computed the posture descriptors p_i for all the videos, we train a multiclass SVM classifier using the pairs (p_i, c_i) composed by the concatenated posture-based descriptor p_i with its corresponding sign class c_i . We follow a one-vs-all strategy and the method from Crammer and Singer [3] to train the system. During testing, given a video captured with kinect, a similar approach is followed to obtain the representation p_i based on posture features. Then, the learned model W is used to compute the response $R_{posture}$ of the input video over the difference sign classes as $R_{posture} = W * p_i$, based purely on hand postures.

3.2. Sign Recognition based on Hand Gestures

This component takes as input RGBD images and the skeleton joints estimated using the method from [1]. The goal of this component is to infer from this skeleton a set of features that enable effective recognition of signs. Towards this goal, from the initial set of 15 3D joints, we only consider a set $J = \{j_1, j_2, \dots, j_{11}\}$ of 11 joints covering the upper body. This is due to the fact that most of the sign languages only use the upper part of the body to define their signs.

Hand Gesture Representation: We propose to represent hand gestures based on relations between the hands and the rest of joints, or parts, of the body. This is motivated by 2 observations: 1) Most sign languages use hands as the main element of the signer. 2) During different hand gestures the hands may follow similar trajectories, however these trajectories can be

defined in the context of different body areas. For example, in Fig. 1, even when the signs in Fig. 1(a) and in Fig. 1(b) have a similar global trajectory, in yellow, the sign in Fig. 1(a) involves hand contact on top of the head, while the sign in Fig. 1(b) involves contact with the lower part of the head.

Given the set J of selected joints where each joint $j = (X, Y, Z)$ is defined by its 3D location. We define the relative body part descriptor (RBPd) as $RBPd = [\delta_1, \delta_2, \dots, \delta_m]$ where $\delta_i = (j_i - j_h)$ is the relative location of each non-hand joint j_i w.r.t. one of the hand joints j_h . We perform this operation for each of the two hands. The final descriptor is defined by the concatenation of the descriptors computed from each hand $RBPd = [RBPd_{right}, RBPd_{left}]$. Note that the user can be at different positions w.r.t. the visual field of the camera and consequently have considerable variation in X, Y and Z coordinates. It is for this reason that building the proposed descriptor, considering relative locations between the hands w.r.t. body, we achieve some level of invariance towards changes in the location of the user. Finally, until now, the estimated input descriptor $RBPd$ constitutes the observation at a specific frame. We extend this frame-level representation to the full gesture sequence by computing this descriptor for each of the n frames of the video $g = [RBPd_1, RBPd_2, \dots, RBPd_n]$.

Visual Dictionary: We build a dictionary of visual words, where each word w_i is derived from RBPds. To this end, we compute all the RBPds from frames of training sequences, z-normalize them, and cluster them using K-means with $K = 95$. This K value was obtained from running the pipeline in the validation set. Each of the means will be the words w_i of the dictionary and each $RBPd$ descriptor will be represented by a word w_i . As a result, each gesture will now be represented by a sequence of words w . This dictionary is stored for the testing stage.

Recognizing signs based on hand gesture features: In this paper, we model the dynamics of the hand gestures using left-right Hidden Markov models (HMMs). Specifically, we train one HMM per sign class. HMMs are a type of statistical model which are characterized by the number of states in the model, the number of distinct observation symbols per state, the state transition probability distribution, the observation symbol probability distribution and the initial state distribution. In our system, the training observations (o_1, o_2, \dots, o_n) are the hand gestures represented as a sequence of words obtained from the visual dictionary. These observations o_i are collected per sign class c_i and used to train each HMM. The state transition probability of each model is initialized with the value 0.5 to allow each state to begin or stay on itself with the same probability. The number of states is different for each model and was determined using the validation data. The number of distinct observation symbols of the models is equal to the number of words from the visual dictionary of gestures ($K = 95$). Furthermore, in order to ensure that

the models begin from their respective first state, the initial state distribution gives all the weight to the first state. Finally, the observation symbol probability distribution matrix of each model is initialized uniformly with the value $1/K$, where K is the number of distinct observation symbols. During training, for each model, the state transition probability distribution, the observation symbol probability distribution and the initial state distribution are readjusted. Once the different HMMs have been trained for each sign class c_i , the system is then ready for sign classification. During testing, given a gesture observation g , sequence of words encoded using the visual dictionary, and a set of pre-trained HMMs Ω , our method selects the class of the model Ω_k that maximizes the likelihood $p(c|g)$ of class c based on gestures features, i.e. $c = \arg \max_{(k)} p(k|g) = \arg \max_{(k)} (\Omega_k(g))$. In this paper we refer to $p(c|g)$ as the sign response $R_{gesture}$ based, purely, on hand gesture features.

3.3. Coupled Sign Language Recognition

For each RGBD sequence, the previous components of the system compute the responses $R_{posture}$ and $R_{gesture}$ over the sign classes based on posture and gesture features, respectively. In order to obtain a final prediction, we define the coupled response R by late fusion of the responses $R_{posture}$ and $R_{gesture}$. To this end, given a set of validation sequences, we compute the responses based on the postures $R_{posture}$ and gestures $R_{gesture}$, and define the coupled descriptor $R = [R_{posture}, R_{gesture}]$ as the concatenation of the two responses. Then, using the coupled descriptors - class label pairs (R_i, c_i) from each validation example we train a multiclass SVM classifier using linear kernels. This effectively learns the optimal linear combination of $R_{posture}$ and $R_{gesture}$. During testing, the sign class \hat{c}_i is obtained as $\hat{c}_i = \arg \max_{(c_k)} (\omega_k \cdot R_i)$. Here, R_i is the coupled response computed from the testing data and $\omega = [\omega_1, \omega_2, \dots, \omega_k]^T$ are the weight vectors from the SVM models.

4. EVALUATION

We evaluate our approach on the ChaLearn Gestures dataset (2013) [6]. We follow a similar procedure as [7, 8, 9] for splitting the data. For the sake of comparison with [7, 8, 9], we report results using as performance metric mean precision, recall and F-Score. Additionally, for the gesture component, we perform experiments on the MSRC-12 dataset [10] following the protocol from [11, 12].

Sign Recognition based on Hand Postures: Fig. 4(a) shows the confusion matrix of the system at recognizing signs based purely on features derived from hand postures. Recognition based on hand posture features was correct in 42% of the cases. This low average is due to the following facts: (1) these signs were captured at a distance around 2-3

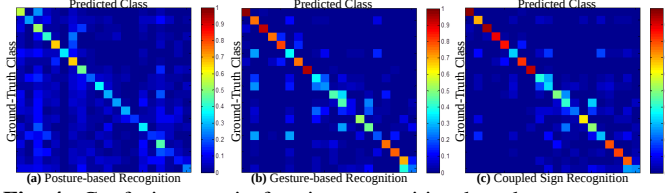


Fig. 4: Confusion matrix for sign recognition based on responses computed from: (a) Hand Postures, (b) Hand Gestures (*RBPD-T*), and (c) late fusion of hand postures and gestures responses

meters from the camera obtaining images with poor resolution, specially for the regions that cover the hands. (2) On many of these signs, the hands come into contact or get very close to the body (see Fig. 1) making it difficult to obtain an accurate segmentation and introducing error in the features computed from the hand region. (3) Some of these signs are defined with very similar sequences of hand postures which leads to miss classification.

Sign Recognition based on Hand Gestures: In this experiment we focus on the recognition of signs based on hand gestures. We evaluate 4 methods to model the gestures: a) the *RBPD* descriptor proposed in section 3.2; b) the *RBPD-T* descriptor which is similar to *RBPD*, however, in this descriptor the relations between the hands and the other parts of the body are estimated taking into account the hand locations in the current frame and the location of the other parts in the next frame. This implicitly adds temporal features to *RPBD*; c) the *HD* descriptor which only considers the location of the hands w.r.t. the torso location; and d) *HD-T*, a time extension of *HD*. The last 2 are methods based on hand trajectories since we only follow the location of the hands over time. Similar to *RBPD*, we trained HMMs (Sec. 3.2) using these methods, *RBPD-T*, *HD*, and *HD-T*, for gesture representation. From these methods, we take the top performing *RBPD-T* for further experiments. Fig. 4(b) shows the confusion matrix of recognizing signs based on hand gesture features.

The first thing to point out is the performance of *HD* which is ~ 21 percentage points (pp) lower than *RBPD*. This suggests that taking into account the different parts of the body indeed encodes more information than methods that only use the global trajectory of the hands. The differences between the performance of *HD* vs *HD-T* and *RBPD* vs *RBPD-T* seem to be minimal. Nevertheless, the temporal extensions seems to bring benefits since there is an improvements of ~ 2.5 pp. As Fig. 4(b) shows there are confusing signs for the gesture module. This is the case for signs with similar gestures which are only different in particular hand postures. In addition, on the Chalearn dataset, we compared our approach against the method from [8] which is based on a combination of HMMs with four skeleton joints from the arms (elbow and wrist locations). This method [8] achieves an improvement of 2 pp over the F-Score of our gestures-based method. On MSRC-12, our method achieves 3 pp over the

Table 1: Gesture-based recognition mean performance.

	ChaLearn dataset [6]			MSRC-12 dataset [10]		
	Precision	Recall	F-Score	Precision	Recall	F-Score
HD	0.33	0.35	0.34	0.76	0.78	0.77
HD-T	0.38	0.36	0.37	0.78	0.78	0.78
RBPD	0.54	0.54	0.54	0.89	0.90	0.89
RBPD-T	0.58	0.58	0.58	0.91	0.92	0.92

Table 2: Comparison with the State of the Art in chronological order. Mean performance over all the 20 sign classes of the ChaLearn 2013 dataset [6].

	Precision	Recall	F-Score
Wu et al. [8]	0.60	0.59	0.60
Yao et al. [9]	-	-	0.56
Pfister et al. [7]	0.61	0.62	0.62
Ours (<i>RBPD-T</i> based)	0.61	0.62	0.62

accuracy reported in [11] which is focused on a feature vector of pairwise joint distances between frames. Also, our method is slightly superior to [12] where they use a more expensive covariance descriptor to relate the body joints.

Coupled Sign Recognition: Here we evaluate the performance of the coupled response from Section 3.3. We report results in Table 2 using as performance metric mean precision, recall and F-Score. As Fig. 4(c) shows, the combination of responses outperforms the overall performance of the method when considering only hand gestures. In addition, the confusion between sign classes is reduced showing the complementarity of both responses (Fig. 4). This is to be expected since some ambiguous cases can be clarified by looking at the relations between parts of the body (Fig. 1(a) vs Fig. 1(b)). Likewise, other ambiguous cases can to be clarified by giving more attention to the hand postures (Fig. 1(b) vs Fig. 1(c)). Finally, compared to [8], our combined method achieves an improvement of ~ 3 pp. Furthermore, our method is still superior by ~ 6 pp over the F-Score performance reported by the method from [9]. This is to be expected since our method explicitly exploits information about hand postures, which [9] ignores. This last feature makes the proposed method more suitable to recognize sign languages where hand posture information is of interest. Even more, our method has a comparable performance to the just-published method from [7] which uses more complex posture descriptors. We cannot report results for the combined method on MSRC-12 [10] since it does not provide the RGBD images, which are required for the postures.

5. CONCLUSION

We presented a method focused on representing each sign by the combination of responses derived from hand postures and hand gestures. Our experiments proved that modeling hand gestures by considering spatio-temporal relations between different parts of the body brings improvements over only considering the global trajectories of the hands. Future work will focus on sign localization/detection.

Acknowledgments: This work is funded by the FWO project G.0.398.11.N.10 “Multi-camera human behavior monitoring and unusual event detection”.

6. REFERENCES

- [1] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, *Real-time human pose recognition in parts from single depth Images*, CVPR, 2011.
- [2] Sergio Belongie and Jitendra Malik, *Matching with Shape Contexts*, IContent-based Access of Image and Video Libraries. Proceedings. IEEE Workshop on, pages 20–26, 2000.
- [3] K. Crammer and Y. Singer, “On the algorithmic implementation of multi-class svms,” in *JMLR*, 2001, vol. 2, pp. 265–292.
- [4] Georgios Th. Papadopoulos, Apostolos Axenopoulos, and Petros Daras, “Real-time skeleton-tracking-based human action recognition using kinect data,” in *Multi-Media Modeling*, 2014.
- [5] Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, and Ming Zhou, *Sign Language Recognition and Translation with Kinect*, FG, 2013.
- [6] S. Escalera, J. Gonzalez, X. Bar, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. J. Escalante, *Multi-modal Gesture Recognition Challenge 2013: Dataset and Results*, ICMI Workshops, 2013.
- [7] T. Pfister, J. Charles, and A. Zisserman, “Domain-adaptive discriminative one-shot learning of gestures,” in *ECCV*, 2014.
- [8] Jiaxiang Wu, Jian Cheng, Chaoyang Zhao, and Hanqing Lu, “Fusing multi-modal features for gesture recognition,” in *ICMI*, 2013.
- [9] Angela Yao, Luc Van Gool, and Pushmeet Kohli, “Gesture recognition portfolios for personalization,” in *CVPR*, 2014.
- [10] Simon Fothergill, Helena M. Mentis, Pushmeet Kohli, and Sebastian Nowozin, “Instructing people for training gestural interactive systems,” in *CHI*, 2012.
- [11] Chris Ellis, Syed Zain Masood, Marshall F. Tappen, Joseph J. Laviola, Jr., and Rahul Sukthankar, “Exploring the trade-off between accuracy and observational latency in action recognition,” in *IJCV*, 2013, vol. 101, pp. 420–436.
- [12] Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban, “Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations,” in *IJCAI*, 2013.